It's Not Just the Prompt: Model Choice Dominates LLM Creative Output

Jennifer Haase
Humboldt University
Berlin, Germany
Weizenbaum Institute
Berlin, Germany
jennifer.haase@hu-berlin.de

Jana Gonnermann-Müller
Interactive Optimization and Learning
Zuse Institute Berlin
Berlin, Germany
Weizenbaum Institute
Berlin, Germany
gonnermann-mueller@zib.de

Paul H. P. Hanel
Department of Psychology
Essex University
Essex, United Kingdom
p.hanel@essex.ac.uk

Nicolas Leins
Interactive Optimization and Learning
Zuse Institute Berlin
Berlin, Germany
leins@zib.de

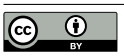
Thomas Kosch
HU Berlin
Berlin, Germany
thomas.kosch@hu-berlin.de

Jan Mendling
Humboldt-Universität zu Berlin
Berlin, Germany
Weizenbaum Institute
Berlin, Germany
jan.mendling@hu-berlin.de

Sebastian Pokutta
Interactive Optimization and Learning
Zuse Institute Berlin
Berlin, Germany
Mathematics and EECS
TU Berlin
Berlin, Germany
pokutta@zib.de

Abstract

Prompt engineering is often treated as a reliable control mechanism for LLM behavior, yet LLM outputs vary even under similar prompts due to stochasticity. We quantify how much output variance is driven by prompt choice versus model choice and by inherent within-LLM stochasticity by evaluating 12 LLMs on 10 creativity prompts in an open-ended divergent-thinking task (AUT), measuring answer quality (originality) and quantity (number of answers), generating 100 samples per prompt. Then, we partition the variance into model, prompt, within-LLM stochasticity, and model×prompt interaction components. Our findings show that model choice is at least as important as prompt choice in this setting. For originality, the model explains 41% of the variance. Prompts explain 36%, and within-model stochasticity explains 11%. For fluency, prompts explain 4% of the variance. Model choice explains 51%, and within-model stochasticity 34%. Beyond variance decomposition, models exhibit persistent "creative fingerprints" in thematic preferences and formatting habits.



CHI EA '26, Barcelona, Spain

ACM ISBN 979-8-4007-2281-3/26/04
https://doi.org/10.1145/3772363.3799284

CCS Concepts

• **Human-centered computing** → **Empirical studies in interaction design**.

Keywords

Large Language Models, Prompt Engineering, Stochasticity, Creativity, Variability, Model Selection

ACM Reference Format:
Jennifer Haase, Jana Gonnermann-Müller, Paul H. P. Hanel, Nicolas Leins, Thomas Kosch, Jan Mendling, and Sebastian Pokutta. 2026. It's Not Just the Prompt: Model Choice Dominates LLM Creative Output. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3772363.3799284

1 Introduction & Background

Prompt engineering has emerged as a central paradigm in HCI work on Large Language Models (LLMs), casting natural language prompts as powerful yet designable inputs. Prior research focuses on how users craft, refine, and structure prompts to achieve specific outcomes (e.g., [17], [1], [15]). Implicit in these efforts is the assumption that prompt choice dominates output behavior and that models respond reliably and consistently to specific prompt formulations.

Yet, LLMs are stochastic systems, and repeated runs with identical prompts can produce strikingly different outputs [4]. Moreover, different models are all trained differently, with different architectures and training data, and therefore exhibit distinct preferences and structural habits that persist regardless of prompt. This inherent variability, both within and between models, has received limited empirical attention. Our study directly interrogates this gap, asking not just *how much* prompts matter, but *what characterizes models* that respond well to different prompting strategies.

Prompt engineering, the design of natural language prompts to steer LLM behavior, is increasingly treated as a user-centered design activity in HCI [8]. Studies have investigated prompt formulation patterns [11], lay users' mental models [17], and interactive refinement tools [1, 5, 15]. Prior work positions prompts as reliable levers that can shape model output, leading to prompt libraries, design guidelines, and tooling ecosystems that treat the prompt as the primary locus of control. Yet LLMs are fundamentally probabilistic: Even with fixed temperature, repeated runs yield divergent outputs [13]. This intrinsic randomness is underexamined in prompting research. While some studies acknowledge inconsistency, few quantify its extent or determine whether it undermines prompt design strategies, mirroring a long-standing critique in psychology. Researchers often treat stimuli as fixed effects, ignoring stimulus-level variability and inflating generalization claims [6, 7]. Prompt engineering research risks the same fallacy by treating prompts as reliable controls while ignoring model and stochasticity through sampling variance [2].

To tackle this challenge, our work quantifies the extent to which variability in LLM creative output is driven by prompt and model choice, as well as within-model stochasticity. We focus on the Alternative Uses Task (AUT), a highly open-ended divergent-thinking task; the balance between model, prompt, and stochastic variance may shift for more constrained tasks. We evaluate 12 LLMs on 10 creativity prompts with 100 samples each (N=12,000) and decompose variance in originality and fluency while examining stable output patterns across models. We find that model choice is at least as important as prompt choice, prompt effects are model-contingent, and models exhibit persistent "creative fingerprints". Prompt-centric user-centered tools should instead prioritize task–model fit and stochasticity-aware workflows, rather than treating prompt optimization as the primary control.

2 Method

This work investigates the relevance of the three types of variance in LLM output for a task in which output variability is not noise but creative idea generation. We ask **RQ1:** *How is variance partitioned across prompt choice, model choice, and stochasticity through sampling?* **RQ2:** *What characterizes models that respond differently to prompting strategies?* We evaluate 12 LLM performances on an open-ended idea generation task on 10 prompts with 100 samples each (N=12,000), decompose variance into model, prompt, interaction, and within-LLM components, and characterize model differences through thematic and structural analysis of 566,916 generated ideas.

2.1 Stimuli: Prompt Design

We designed 10 prompts for the Alternative Uses Task (AUT) [3], a standard creativity assessment where participants generate novel uses for common objects (here: a plastic bottle). Prompts span six strategic categories (P1–P6) [10] plus four minor variations (P7–P10) [16] to test robustness. Table 1 summarizes the strategies.

Table 1: Prompting Strategies (P1–P10).

ID	Strategy	Key Manipulation
P1	Direct Instruction	Baseline AUT task: "Think of all kinds of things you could do with a plastic bottle."
P2	One-Shot Example	Adds concrete example: "A plastic bottle can be used to store yarn without tangling."
P3	Heuristic/ Domain	Cross-domain cues: "Think across domains: art, survival, education, sci-fi..."
P4	Anticipatory	Avoidance framing: "Avoid generic ideas like holding water..."
P5	Zero-Shot CoT	Reasoning scaffold: "Think step by step: What are the physical features?"
P6	Creative Persona	Identity priming: "You are the most creative person on the planet."
P7	Phrasing Variation	Synonymous rewording of P1
P8	Format Constraint	Output restriction: "no titles or colons"
P9	Information Order	Same content as P1, reordered
P10	Typo Injection	P1 with spelling/syntax errors

2.2 Models and Procedure

We tested 12 LLMs spanning proprietary and open-source systems: Claude Sonnet 4.5, DeepSeek V3.2, GPT 5.1, GPT 5.2, GPT OSS 120B, Gemini 3 Pro, Gemma 3 27B, Grok 4.1, Llama 3.3 70B, Mistral Nemo, Qwen 3 235B, and Qwen 3 235B Thinking. Proprietary models were accessed via official APIs; open-weight models were hosted locally using Ollama. We employed *default* hyperparameters as defined by each provider, since temperature effects vary across architectures [14] and defaults maximize ecological validity. For each model × prompt pair, we generated 100 independent outputs, totaling 12,000 generations. Data were collected in December 2025; 11,870 (98.9%) were successfully scored using the OCSAI scoring system [12], yielding 566,916 individual ideas.

2.3 Evaluation

Creative Performance Variance Decomposition. We evaluated outputs on two metrics: *originality* (scored 1–5 via OCSAI [12]; response-level scores computed as mean across all ideas) and *fluency* (count of valid ideas per response after filtering empty strings and formatting artifacts). We partition total variance using a two-way crossed random-effects model:

$$Y_{mpr} = \mu + \alpha_m + \beta_p + (\alpha\beta)_{mp} + \epsilon_{mpr} \quad (1)$$

where μ is the grand mean, $\alpha_m \sim N(0, \sigma^2_{\text{model}})$ captures systematic differences between LLMs, $\beta_p \sim N(0, \sigma^2_{\text{prompt}})$ captures prompt effects, $(\alpha\beta)_{mp}$ their interaction (model-specific prompt sensitivity), and $\epsilon_{mpr} \sim N(0, \sigma^2_{\text{within}})$ captures within-LLM stochasticity through sampling variance. Parameters were estimated via Restricted Maximum Likelihood (REML). We report Intra-class Correlation Coefficients (ICC) as variance proportions, with 95% confidence intervals computed via stratified bootstrap (1,000 iterations).

Model Characterization. To understand *how* models differ beyond aggregate scores, we analyzed idea content on two dimensions: (1) *Thematic categorization*: We developed a keyword-based taxonomy with 105 keywords mapped to 10 functional domains (e.g., *Gardening*, *Survival*, *Art*). Ideas were classified via deterministic substring matching. We computed *thematic entropy* (Shannon entropy of category distribution) to quantify breadth vs. depth of each model's creative range [9]. (2) *Structural profiling*: We extracted formatting features: punctuation style (colons vs. dashes), verbosity (character count), and numeric density, to identify "alignment fingerprints" that persist across prompts.

3 Results

Variance Decomposition. Our analysis reveals that prompts matter—but not uniformly (Figure 1). For originality (idea quality), prompt strategy explains 36% of variance, comparable to model choice (41%). Within-LLM stochastic variance accounts for 11%, with 12% from model×prompt interactions. For fluency (number of ideas), prompts barely matter: only 4% of variance explained. Model choice dominates (51%), followed by within-LLM stochasticity through sampling variance (34%). How many ideas an LLM generates is largely intrinsic to its architecture, at least when not specifically prompted for.

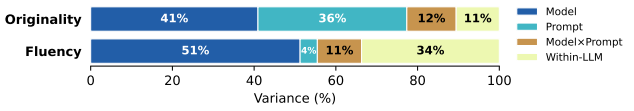


Figure 1: Variance Decomposition. Prompts explain 36% of originality variance but only 4% of fluency.

Prompt Strategies. Not all prompt variations are equally informative (Figure 2). *Discriminative prompts* like the Persona prompt (P6: "You are the most creative person...") act as stress tests, yielding highest between-model variance (0.112 vs. baseline 0.037) and cleanly separating capable models from less original ones. Grok 4.1 and Gemma 3 27B show massive improvements under P6 (Cohen's $d > 2.0$), while smaller models show negligible gains. Heuristic (P3) and Anticipatory (P4) prompts similarly elicited higher originality with higher between-model variance. *Constraining prompts* like Format Constraint (P8) reveal unintended effects: structural constraints can cause "semantic collapse", reducing idea uniqueness by 5.9 percentage points despite near-perfect compliance. *Non-discriminative prompts* like One-Shot Example (P2) and Chain-of-Thought (P5) did not improve originality relative to baseline. These strategies may not generalize to creative tasks.

Model Trade-offs. The large model effect (41–51% of variance) reflects fundamental architectural trade-offs. Gemini 3 Pro achieves highest mean originality (4.06/5) by generating fewest ideas (avg. 13.7). The Qwen 3 Thinking variant improves originality by 3.4% over base Qwen and reduces fluency by 38%. In contrast, GPT 5.1 and GPT 5.2 generate >150 ideas per prompt but with lower mean originality. Grok 4.1 produced 5.5× more top-1% originality responses than random chance, despite ranking only 3rd in mean, suggesting it excels at occasional best-performance rather than consistent performance. To control for verbosity bias (idea length correlates with originality scores, $r \approx 0.1$–0.4), we computed length-adjusted scores: model rankings proved robust, with Gemini 3 Pro retaining #1 and Qwen 3 Thinking improving from #6 to #3, as the most concise yet original model.

Creative Fingerprints. Beyond aggregate variance, models exhibit distinct "creative personalities" that persist across prompts (Figure 3). Clustering 566,916 ideas into 10 functional categories reveals stable thematic specializations: Mistral Nemo concentrates 31% of ideas into "Gardening"; Gemini 3 Pro over-indexes on Art/Decor (24.5%); Llama 3.3 concentrates on Storage/Organization (20.5%); Grok 4.1 shows unusual strength in Music/Sound (6.1%, 2× average). DeepSeek Reasoner distributes ideas broadly (highest thematic entropy: 2.16). Models also adhere to rigid formatting habits: Gemini includes numerical measurements in 11.8% of ideas; GPT 5.1 uses dash-based structures; Claude favors colon-separated titles. When P8 forbids a model's native format ("no titles or colons"), idea uniqueness drops 5.9 percentage points vs. baseline for Claude, which we interpret as a "structural collapse" effect. The models with the highest compliance with syntactic constraints inadvertently restrict semantic diversity, producing over 100 exact duplicate phrasings across models. These results show that some prompts can come with very unintended side effects.

4 Discussion

Our findings challenge the prompt-centric view of LLM interaction. While prompts matter for steering *quality* (36% of originality variance), model choice is similarly important (41%), and model×prompt interactions (12%) reveal that prompt effectiveness is strongly model-contingent: the "best" prompt depends on which model you use. While model dominance may partly reflect capability gaps between small and large models, the creative fingerprint analysis shows that models differ not only in level but in consistent, measurable patterns, like thematic preferences and formatting habits that persist across prompts. This suggests that model selection deserves equal attention to prompt design in creative AI tools. Current interfaces typically treat model choice as a background setting; our results argue for foregrounding it as a primary control. Our analysis reveals three dimensions of model×prompt fit: (1) *Generalist vs. Specialist*: high-entropy models (DeepSeek, Qwen) explore broadly while low-entropy models (Mistral, Llama) concentrate in specific domains; (2) *Quality vs. Quantity*: reasoning-oriented models (Gemini, Qwen Thinking) produce fewer but more original ideas while high-fluency models (GPT 5.x) maximize number of ideas generated; (3) *Prompt Sensitivity*: some models (Grok, Gemma) show large gains under Persona prompts while others are relatively unresponsive. These dimensions suggest that "good" fit is task-dependent:

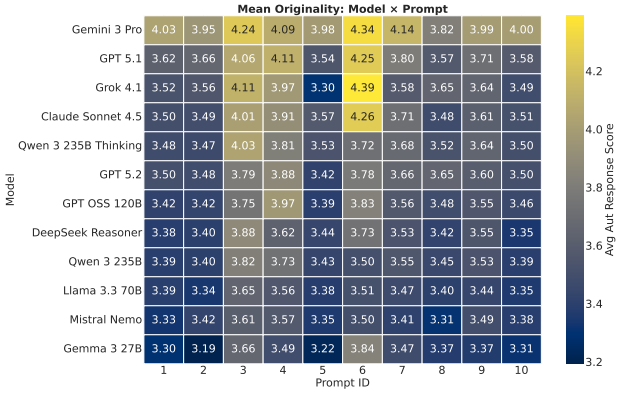


Figure 2: Model×Prompt Interaction. Mean originality scores reveal prompt effectiveness is model-contingent.

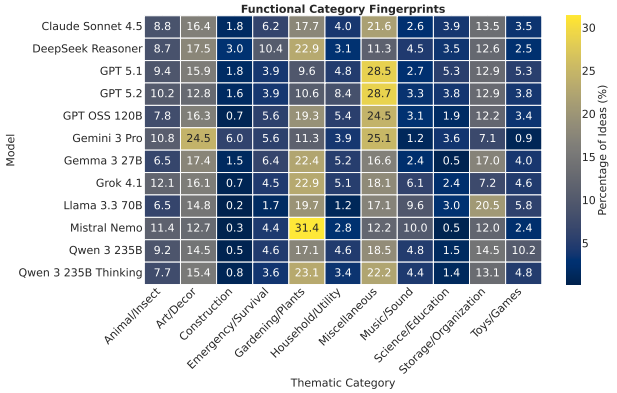


Figure 3: Functional Category Fingerprints. Each cell shows the percentage of a model's ideas falling into each thematic category.

exhaustive brainstorming benefits from high-fluency generalists; targeted ideation may benefit from specialists paired with discriminative prompts. For creative AI interfaces, these results suggest foregrounding model choice via lightweight cues about model "personalities" (thematic strengths and quality–quantity trade-offs), treating regeneration as a first-class interaction given substantial within-model variance, and curating prompt libraries as model-specific rather than one-size-fits-all.

Future Research. Future work should test whether model-aware interaction improves creative performance and decision-making compared to prompt-centric workflows. We propose a controlled user study in which participants perform creative ideation tasks under three interface conditions: (1) a standard prompt-only interface, (2) an interface that foregrounds model choice without explanation, and (3) a model-aware interface that exposes concise model fingerprints (e.g., quality–quantity trade-offs, thematic strengths, and prompt sensitivity) and supports regeneration as an explicit strategy. Outcomes would include creative quality, exploration breadth, user confidence, and strategy adaptation over repeated trials. This design directly probes the core implication of our findings: whether making model differences legible enables users to better navigate stochasticity through sampling, rather than treating variability as noise to be eliminated. By focusing on selection and interaction strategies rather than prompt optimization, such a study would establish whether model awareness constitutes a genuine usability and creativity advantage, and whether the field should rethink the prompt as the primary locus of control in LLM-based tools. The dominance of model and stochastic variance underscores the need to better analyze and understand their roles in AI-supported tasks. Future work should generalize variance-aware designs beyond creativity to domains like writing, coding, and reasoning, where prompt engineering is similarly assumed to be the primary control mechanism. We also call for research on how to make model fingerprints legible to users without large-scale experiments, e.g., through community benchmarks, provider documentation, or lightweight diagnostic prompts, since discovering model personalities currently requires resources beyond typical users or design teams.

Limitations. This study is *exploratory* in the way that it identifies dimensions of model difference but does not provide validated selection guidelines. We studied one creative task (AUT), a highly open-ended divergent-thinking task; the model–prompt variance balance may shift substantially for more constrained tasks (e.g., code generation, summarization) where strict constraint satisfaction matters more than divergence. AUT may partly reflect retrieval rather than creativity; we treat it as a standard creative task, but future work should test more complex tasks. The originality metric relies on automated scoring trained on human-generated responses; a brief discussion of whether OCSAI generalizes to LLM-generated text, or a small-scale human validation, would strengthen confidence in the findings. Our "model" factor conflates architecture, parameter count (27B vs. 235B+), and RLHF tuning; the dominance of model variance is partly expected given these capability gaps. Disentangling these requires controlled experiments on open-weight models. Model fingerprints may drift as providers update systems. Whether users perceive and benefit from these patterns warrants future investigation.

Conclusion. We decomposed variance across 12,000 LLM creative outputs in an open-ended divergent-thinking task, finding that while prompts explain 36% of originality variance, model choice is similarly important (41%) and interacts strongly with prompt effects (12%). Beyond variance, we characterized *how* models differ: along generalist–specialist, quality–quantity, and prompt-sensitivity dimensions. This exploratory analysis suggests that model selection deserves equal attention to prompt design in creative AI research and tooling. The question is not just "which prompt?" but "which model, for which task, with which prompt?" We offer this work as groundwork for a research agenda on model×prompt fit in open-ended creative tasks, aiming to understand when and why specific model–prompt combinations succeed.

Acknowledgments

This work was partially supported by the Federal Ministry of Research, Technology and Space (grant 16DII133; originally funded by the German Federal Ministry of Education and Research in

2017), the Deutsche Forschungsgemeinschaft (DFG) through grants 496119880 (VisualMine) and 531115272 (ProImpact), and the DFG Cluster of Excellence MATH+ (EXC-2046/1, project ID 390685689). Additional support was provided through CRC FONDA (project ID 414984028) and by the Zuse Institute Berlin via RISE@ZIB services for hosting the LLM models.

References

[1] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3613904.3642016

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922

[3] Paul R. Christensen, Joy P. Guilford, and Robert C. Wilson. 1960. *Alternate Uses Test.* Sheridan Psychological Services, Beverly Hills, CA.

[4] Jennifer Haase, Paul H. P. Hanel, and Sebastian Pokutta. 2025. Has the Creativity of Large-Language Models Peaked?: An Analysis of Inter- and Intra-LLM Variability. *Journal of Creativity* (Nov. 2025), 100113. doi:10.1016/j.yjoc.2025.100113

[5] Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. 2022. PromptMaker: Prompt-based Prototyping with Large Language Models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 35, 8 pages. doi:10.1145/3491101.3503564

[6] Charles M. Judd, Jacob Westfall, and David A. Kenny. 2012. Treating Stimuli as a Random Factor in Social Psychology: A New and Comprehensive Solution to a Pervasive but Largely Ignored Problem. *Journal of Personality and Social Psychology* 103, 1 (2012), 54–69. doi:10.1037/a0028347

[7] Charles M. Judd, Jacob Westfall, and David A. Kenny. 2017. Experiments with More Than One Random Factor: Designs, Analytic Models, and Statistical Power. *Annual Review of Psychology* 68, 1 (Jan. 2017), 601–625. doi:10.1146/annurev-psych-122414-033702

[8] Tanya Kraljic and Michal Lahav. 2024. From Prompt Engineering to Collaborating: A Human-Centered Approach to AI Interfaces. *Interactions* 31, 3 (May 2024), 30–35. doi:10.1145/3652622

[9] J. Lin. 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory* 37, 1 (Jan. 1991), 145–151. doi:10.1109/18.61115

[10] Xiaoqiang Lin, Zhaoxuan Wu, Zhongxiang Dai, Wenyang Hu, Yao Shu, See-Kiong Ng, Patrick Jaillet, and Bryan Kian Hsiang Low. 2024. Use Your INSTINCT: INSTruction Optimization for LLMs usIng Neural Bandits Coupled with Transformers. arXiv:2310.02905 [cs] doi:10.48550/arXiv.2310.02905

[11] Vivian Liu and Lydia B Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–23. doi:10.1145/3491102.3501825

[12] Peter Organisciak, Denis Dumas, and Paul J. Silvia. 2025. Open Creativity Scoring: A Validated and Open-Source System for Evaluating Divergent Thinking. *Creativity Research Journal* 37, 1 (2025), 1–22. doi:10.1080/10400419.2024.2392749

[13] Andrei Paleyes, Radzim Sendyka, Diana Robinson, Christian Cabrera, and Neil D. Lawrence. 2025. Prompt Variability Effects On LLM Code Generation. arXiv:2506.10204 [cs] doi:10.48550/arXiv.2506.10204

[14] Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Temperature is a Creativity Parameter for Language Models. *Proceedings of the 15th International Conference on Computational Creativity* (2024), 240–249.

[15] Savvas Petridis, Ben Wedin, Ann Yuan, James Wexler, and Nithum Thain. 2024. ConstitutionalExperts: Training a Mixture of Principle-based Prompts. arXiv:2403.04894 [cs] doi:10.48550/arXiv.2403.04894

[16] Abel Salinas and Fred Morstatter. 2024. The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance. arXiv:2401.03729 [cs.CL] doi:10.48550/arXiv.2401.03729

[17] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–21. doi:10.1145/3544548.3581388